WHAT IS CLAIMED IS:

1. A method of determining a statistical model for predicting disease risk for a member of a population,

   a. collecting a plurality of sets of data, each of said sets of data associated with one member of said population, and comprising data of a first type, data of a second type, and an indicator of disease status of said one member associated with said set;

   b. selecting a candidate statistical model for calculating said disease risk as a function of data of said first type, said candidate model dependent on a plurality of parameters;

   c. determining a plurality of weights, each one of said weights associated with one of said sets of data and indicating a statistical significance of said one of said sets of data, wherein weights associated with sets of said data having like data of said second type are the same; and

   d. optimizing said parameters of said candidate model by fitting said plurality of sets of data to said candidate model, taking into account said weights.

2. The method of claim 1, wherein data of said first type is non-genetic data and data of said second type is genetic data.

3. The method of claim 1, wherein said corresponding weights are used to assess a goodness of said fitting.

4. The method of claim 1, wherein said determining comprises:

   a. grouping said collected data into groups such that all sets of data within each said group have like data of said second type, one of said groups being a reference group which contains sets of data having data of said second type like data of said second type obtained from said member of said population; and

   b. determining a group weight for each said group, whereby said group weight is the corresponding weight for each set of data within said each group.

5. The method of claim 4, wherein the group weight of said reference group has a value of one and each of the other group weights has a value between zero and one.

6. The method of claim 5, wherein said other group weights are optimized by minimizing a target function, said target function dependent on a plurality of residuals, one of said residuals for each of the data sets in said reference group.

7. The method of claim 6, wherein a residual for the *i*th one of said data sets is the difference between the value of the indicator of disease status contained in said *i*th data set and the value of disease risk for the member associated with said *i*th data set , said value of disease risk calculated from said candidate model with said parameters optimized for a given set of group weights by fitting data sets in groups other than the reference group to said candidate model.

8. The method of claim 7, wherein said target function is of the form:

$$f = \sum w_i \phi(r_i),$$

where

w_i is the corresponding weight for data set i; and

r_i is the residual for data set i.

9. The method of claim 1, wherein data of said first type comprises data indicative of time.

10. The method of claim 9, wherein said candidate model is a Cox proportional hazard regression model.

11. The method of claim 9, wherein said candidate model is a disease risk function of the form:

$$R(t) = 1 - \exp\{-\int_0^t h(u)du\},$$

where

$R(t)$ represents said disease risk at a given time t;

$h(u)$ is of the form:

$$h(u) = h_0(u)\exp(\sum_{1}^{n_c} \beta_i x_i);$$

$h_0(u)$ is dependent only on u;

x_i is a variable indicative of a disease risk factor, said collected data containing a plurality of values of x_i;

β_i is a coefficient for x_i; and

n_c is the number of coefficients in said disease risk function.

12. The method of claim 1, wherein said collecting comprises imputing missing data to said plurality of data sets.

13. The method of claim 1, wherein each corresponding weight is weighted by an adjustment factor indicative of the representativeness of the member associated with said each corresponding weight.

14. The method of claim 13, wherein an adjustment factor a_i for a data set obtained from a member *i* of said population is calculated as:

$$a_i = \frac{n_i^p}{n_i^s},$$

where n_i^p is the number of members in said population who share a same set of characteristics with said member *i*, and n_i^s is the number of members associated with said collected data who share said set of characteristics.

15. The method of claim 14, wherein said set of characteristics comprises non-genetic factors.

16. The method of claim 14, wherein said set of characteristics comprises genetic factors.

17. The method of claim 14, wherein said set of characteristics comprises both genetic and non-genetic factors.

18. The method of claim 14, wherein said set of characteristics are selected from the group of age, gender, race, body mass index, smoking status, hypertension, cholesterol level, personal health history, and family health history.

19. The method of claim 1, comprising calculating a disease risk for said member of said population with said disease risk prediction model.

20. A computing system adapted for perform the method of any one of claims 1 to 19.

21. An article of manufacture comprising

a computer readable medium embedded thereon computer executable instructions, which when executed by a computer causes said computer to determine a statistical model for predicting disease risk for a member of a population by

a. collecting a plurality of sets of data, each of said sets of data associated with one member of said population, and comprising data of a first type, data of a second type, and an indicator of disease status of said one member associated with said set;

b. selecting a candidate statistical model for calculating said disease risk as a function of data of said first type, said candidate model dependent on a plurality of parameters;

c. determining a plurality of weights, each one of said weights associated with one of said sets of data and indicating a statistical significance of said one of said sets of data, wherein weights associated with sets of said data having like data of said second type are the same; and

d. optimizing said parameters of said candidate model by fitting said plurality of sets of data to said candidate model, taking into account said weights.

22. A method of imputing missing data indicative of a plurality of factors, comprising:

a. determining a correlation between said plurality of factors;

b. grouping said factors into batches such that all factors in each said batch are correlated; and

c. imputing missing data for factors in one said batch at a time.

23. A method of grouping a plurality of data sets into groups, comprising:

a. dividing said plurality of data sets into two or more groups depending on data indicative of a factor of a first type in each of said data sets;

b. determining if a criterion is met after said dividing, said criterion is evaluated based on data of a second type in each of said data sets; and

c. when said criterion is not met, regrouping said plurality of data sets back into one group.

24. The method of claim 23, wherein said dividing is performed recursively on each group of a division.

25. The method of claim 24, wherein divisions at different levels are made dependent on data indicative of different factors.

26. The method of claim 25, wherein a branch of said recursive division is terminated at the level at which said criterion is not met.

27. A method of weighing a plurality of data sets, each one of said data sets associated with a member of a population, comprising:

weighing each set of said plurality of data sets by a weight indicative of the representativeness of the member associated with said each set, wherein a weight a_i for a data set obtained from a member *i* of said population is calculated as:

$$a_i = \frac{n_i^p}{n_i^s},$$

where n_i^p is the number of members in said population who share a same set of characteristics with said member *i*, and n_i^s is the number of members associated with said collected data who share said set of characteristics.